

Teachers developing assessment for learning: impact on student achievement

Dylan Wiliam^{1*}, Clare Lee², Christine Harrison³ & Paul Black³

¹Educational Testing Service, NJ, USA; ²Warwickshire County Council, UK;

³King's College London, UK

While it is generally acknowledged that increased use of formative assessment (or assessment for learning) leads to higher quality learning, it is often claimed that the pressure in schools to improve the results achieved by students in externally-set tests and examinations precludes its use. This paper reports on the achievement of secondary school students who worked in classrooms where teachers made time to develop formative assessment strategies. A total of 24 teachers (2 science and 2 mathematics teachers, in each of six schools in two LEAs) were supported over a six-month period in exploring and planning their approach to formative assessment, and then, beginning in September 1999, the teachers put these plans into action with selected classes. In order to compute effect sizes, a measure of prior attainment and at least one comparison group was established for each class (typically either an equivalent class taught in the previous year by the same teacher, or a parallel class taught by another teacher). The mean effect size in favour of the intervention was 0.32.

Introduction

Reviews of research by Natriello (1987) and Crooks (1988) and more recently by Black and Wiliam (1998a) have demonstrated that substantial learning gains are possible when teachers introduce formative assessment into their classroom practice. It is also clear from these reviews, and from other studies (see Black & Atkin, 1996) that achieving this is by no means straightforward. As Black and Wiliam (1998b) point out, these changes are hard to implement even in ideal conditions:

Thus the improvement of formative assessment cannot be a simple matter. There is no 'quick fix' that can be added to existing practice with promise of rapid reward. On the contrary, if the substantial rewards of which the evidence holds out promise are to be secured, this will only come about if each teacher finds his or her own ways of incorporating the lessons and ideas that are set out above into her or his own patterns of classroom work. This can only happen relatively slowly, and through sustained programmes of professional development and support. This does not weaken the message here—indeed, it should be a sign of its authenticity, for lasting and fundamental improvements in teaching and learning can only happen in this way. (p. 15, emphasis in original)

*Corresponding author: Learning and Teaching Research Center, ETS, Rosedale Road (ms 04-R), Princeton, NJ 08541, USA. Email: dwiliam@ets.org

However, the introduction of high-stakes state-mandated testing, such as now exists in England and most states in the USA, makes the effective implementation of formative assessment even more difficult. This is because, although work by Nuthall and Alton-Lee (1995) has shown that teaching for understanding, rather than rote recall, results in better long-term retention, attempts to maximize student and school scores appear to result in a lack of attention to the kinds of higher-order thinking involved in formative assessment (Paris *et al.*, 1991). Indeed, it appears as if there is a widespread belief that teaching well is incompatible with raising test scores.

There have been some studies that have shown that the use of higher-order goals is compatible with success, even when attainment is measured in such narrow terms as scores on external tests. In a three-year study of schools in the mid-west of the USA, Newmann *et al.* (2001) found that students whose teachers used authentic classroom tasks (defined as requiring construction, rather than reproduction of knowledge, disciplined inquiry, and value beyond school) out-performed students not given such work, and that the size of the effects (as measured by standardized effect size) was substantial. In reading, writing and mathematics, the standardized effect sizes were 0.43, 0.52 and 0.64 respectively, with significant aptitude-treatment interactions favouring high-achievers in reading and low-achievers in mathematics.

In another three-year study of two secondary (11–16) schools in England, Boaler (2002) compared two schools. One school (Phoenix Park) used a ‘reform’ approach to the teaching of mathematics, emphasizing higher-order thinking, and students’ responsibility for their own learning, while the other (Amber Hill) used a ‘traditional’ approach emphasizing practice of test items. Although matched in terms of prior achievement, students at Phoenix Park outperformed those at Amber Hill in the national school-leaving examination (the General Certificate of Secondary Education, or GCSE) by, on average, one third of a grade, equivalent to a standardized effect size of 0.21.

These studies are useful in pointing out that attention to higher-order goals in teaching can result in higher attainment, even when such attainment is measured principally in terms of lower-order goals. However, since these studies were not based on direct experiments, there is always the possibility that, in Newmann *et al.*’s (2001) study, the teachers using more authentic activities were just better teachers, and that the choice of authentic activities was incidental to their success. Similarly, in Boaler’s (2002) study, it could be that the teachers teaching at Phoenix Park were just better teachers, drawn to the school by its progressive ethos.

In order to draw clear policy implications regarding the utility of formative assessment, we therefore decided that it was necessary to undertake a more direct experiment, in which the confounding of variables, while not being entirely removed, was reduced, by asking teachers to incorporate formative assessment (or assessment for learning as it is sometimes called) into their classroom practice, and comparing the performance of their students with those of other classes at the same school. This work was undertaken in the King’s-Medway-Oxfordshire Formative Assessment Project (KMOFAP), funded initially by the Nuffield Foundation (as the *Developing Classroom Practice in Formative Assessment* project) and subsequently by

the United States National Science Foundation through their support of our partnership with the Stanford Classroom Assessment Project to Improve Teaching And Learning (CAPITAL; NSF Grant REC-9909370).

Research strategy

The central tenet of the research project was that if the promise of formative assessment was to be realized, traditional research designs—in which teachers are ‘told’ what to do by researchers—would not be appropriate. This is not because teachers somehow fail accurately to put into practice the prescriptions of researchers, but because the general principles emerging from the research underdetermine action—put simply, they do not tell you what to do.

Teachers will not take up attractive sounding ideas, albeit based on extensive research, if these are presented as general principles which leave entirely to them the task of translating them into everyday practice—their classroom lives are too busy and too fragile for this to be possible for all but an outstanding few. What they need is a variety of living examples of implementation, by teachers with whom they can identify and from whom they can both derive conviction and confidence that they can do better, and see concrete examples of what doing better means in practice. (Black & Wiliam, 1998b, pp. 15–16)

This difficulty of ‘putting research into practice’ is not the fault of the teacher. But nor is it a failing in the research. Because our understanding of the theoretical principles underlying successful classroom action is weak, research can never tell teachers what to do. Indeed, given the complexity of classrooms, it seems likely that the positivist dream of an effective theory of teacher action—which would spell out the ‘best’ course of action given certain conditions—is not just difficult and a long way off, but impossible in principle (Wiliam, 2003).

For these reasons we decided that we had to work in a genuinely collaborative way with a small group of teachers, suggesting directions that might be fruitful to explore, and supporting them as well as we could, but avoiding the trap of dispensing ‘tips for teachers’. At first, it seems likely that the teachers did not believe this. They seemed to believe that the researchers were operating with a perverted model of discovery learning in which the researchers knew full well what they wanted the teachers to do, but didn’t tell them, because they wanted the teachers ‘to discover it for themselves’. However, after a while, it became clear that there was no prescribed model of effective classroom action, and each teacher would need to find their own way of implementing these general principles in their own classrooms.

The sample

We began by selecting two local education authorities (LEAs) where we knew there was support from the authority for attempting to develop formative assessment, and, just as importantly, where there was an individual officer who could act as a link between the research team and the schools, and provide a local contact for ad hoc support for the teachers. In this regard, we are very grateful to Sue Swaffield from

Table 1. The six schools involved in the project

School	Abbreviation	Description
Brownfields	BF	Boys
Century Island	CI	Mixed
Cornbury Estate	CE	Mixed
Riverside	RS	Mixed
Two Bishops	TB	Mixed
Waterford	WF	Girls

Medway and Dorothy Kavanagh from Oxfordshire who, on behalf of their authorities, helped to create and nurture our links with the schools. Their involvement in both planning and delivering the formal in-service sessions, and their support ‘on the ground’ have been invaluable, and it is certain that the project would not have been as successful without their contributions.

Having identified the two authorities, we asked each authority to select three schools that they felt would be suitable participants in the project. We were very clear that we were not looking for ‘representative’ or typical schools. From our experiences in curriculum development—for example in graded assessment (Brown, 1988)—we were aware that development is very different from implementation. What we needed were schools that had already begun to think about developing assessment for learning, so that with these teachers we could begin to produce the ‘living examples’ alluded to earlier to use in further dissemination.

Each authority identified three schools that were interested in exploring further the possibility of their involvement, and three of us (PB, CH and DW) visited each school with the LEA officer to discuss the project with the head teacher and other members of the senior management team. All six schools identified agreed to be involved. Brief details of the six schools are shown in Table 1 (the names of all schools and teachers are, of course, pseudonyms).

In our original proposal to the Nuffield Foundation, we had proposed to work only with mathematics and science teachers, partly because of our greater expertise in these subjects, but also because we believed that the implications for assessment for learning were clearer in these areas. In order to avoid the possible dangers of isolation, our design called for two mathematics and two science teachers at each school to be involved.

The choice of teachers was left to the school, and a variety of methods was used. In some schools, the heads nominated a head of department together with a teacher in their first or second year of teaching. In another school, in order to ensure a commitment to the project, the head teacher insisted that both the heads and deputies of the mathematics and science departments were involved. In other schools, teachers appeared to be selected because, in the words of one head, ‘they could do with a bit of inset’. In the event, while our schools were not designed to be representative, there was a considerable range of expertise and experience amongst the 24 teachers selected—five of the teachers were heads of department,

Table 2. Pattern of in-service sessions held

INSET	Held		Format	Focus
A	February	1999	whole-day, London	introduction
B	May	1999	whole-day, London	developing action plans
C	June	1999	whole-day, London	reviewing and revising action plans
	September	1999	half-day, LEA based	reviewing and revising action plans
D	November	1999	whole-day, London	sharing experiences, refining action plans, planning dissemination
E	January	2000	whole-day, London	research methods, dissemination, optional sessions including theories of learning
F	April	2000	whole-day, London	integrating learning goals with target setting and planning, writing personal diaries
G	June	2000	whole-day, London	action plans and school dissemination plans, data analysis 'while you wait'

five were deputy heads of department and the remaining 14 occupied a range of positions within their schools, mostly at a relatively junior level.

The intervention

The intervention had two main components:

- a series of half-day and one-day in-service sessions, during which teachers would be introduced to our view of the principles underlying formative assessment, and have a chance to develop their own plans;
- visits to the schools, during which the teachers would be observed teaching by project staff, have an opportunity to discuss their ideas, and plan how they could be put into practice more effectively.

In our original proposal, we had envisaged a series of nine half-day in-service sessions, some, involving all the teachers, to be held in London, and others conducted in the LEA in order to reduce the teachers' travelling time. In the event, only one such LEA-based session was held, because the teachers felt that they gained a great deal from working with teachers in the other authority. As a result, since most of the teachers would spend two or three hours travelling to reach London, all the inset sessions, apart from the one LEA-based session, took the form of full-day sessions (typically 10 a.m. to 4 p.m.). Although we paid the costs of travel and replacement teaching, since not all teachers could attend all the insets, the savings allowed a total of six-and-a-half days' inset (rather than the proposed four-and-a-half).

The pattern of insets is shown in Table 2 (subsequent insets were held as part of

Table 3. Frequencies of activities in the action plans of 24 teachers

Category	Activity	Frequency
Questioning	Teacher questioning	11
	Pupils writing questions	8
	Existing assessment: pre-tests	4
	Pupils asking questions	4
Feedback	Comment-only marking	6
	Existing assessment: re-timing	4
	Group work: test review	4
Sharing criteria with learners	Course work: marking criteria	5
	Course work: examples	4
	Start of lesson: making aim clear	4
	Start of lesson: setting targets	1
	End of lesson: teacher's review	1
	End of lesson: pupils' review	4
	Group work: explanation	2
	Involving classroom assessment	2
Self-assessment	Self-assessment: traffic lights	11
	Self-assessment: targets	5
	Group work: test review	6
	Self-assessment: other	7
	Pupil peer-assessment	5
	Group work: revision	1
General	Including parents	1
	Posters	1
	Presentations	1
Total		102

the NSF-funded work on the CAPITAL project, but the data reported here relate to the original project, from January 1999 to August 2000.

The key feature of the inset sessions was the development of action plans. Since we were aware from other studies that effective implementation of formative assessment requires teachers to re-negotiate the 'learning contract' (c.f. Brousseau, 1984) that they had evolved with their students, we decided that implementing formative assessment would best be done at the beginning of a new school year. For the first six months of the project, therefore, we encouraged the teachers to experiment with some of the strategies and techniques suggested by the research, such as rich questioning, comment-only marking, sharing criteria with learners, and student peer-assessment and self-assessment. Each teacher was then asked to draw up, and later to refine, an action plan specifying which aspects of formative assessment they wished to develop in their practice and to identify a focal class with whom these strategies would be introduced in September 1999. Although there was no inherent structure in these plans (see below), the teachers being free to explore whatever they wished, we did find that they could be organized under the broad headings shown in Table 3. In all, the 24 teachers included a total of 102 activities in their action plans—an average of just over four each.

Most of the teachers' plans contained reference to two or three important areas in their teaching where they were seeking to increase their use of formative assessment, generally followed by details of strategies that would be used to make this happen. In almost all cases the plan was given in some detail, although many teachers used phrases whose meanings differed from teacher to teacher (even within the same school).

Practically every plan contained some reference to focusing on or improving the teacher's own questioning techniques although only 11 gave details on how they were going to do this (for example, using more open questions, allowing students more time to think of answers or starting the lesson with a focal question). Others were less precise (for example, using more sustained questioning of individuals, or improving questioning techniques in general). Some teachers mentioned planning and recording their questions. Many teachers also mentioned involving students more in setting questions (for homework, or for each other in class). Some teachers also saw existing national curriculum tests as a source of good questions.

Using comment-only marking was specifically mentioned by nearly half the teachers, although only 6 of the teachers included it as a specific element in their action plans. Some of the teachers wanted to reduce the use of marks and grades, but foresaw problems with this, given school policies on marking of student work. Four teachers planned for a module test to be taken well before the end of the module thus providing time for remediation.

Sharing the objectives of lessons or topics was mentioned by most of the teachers, through a variety of techniques (using a question that the students should be able to answer at the end of the lesson, stating the objectives clearly at the start of the lesson, getting the students to round up the lesson with an account of what they had learned). About half the plans included references to helping the students understand the marking criteria used for investigative or exploratory work, generally using exemplars from students from previous years. Exemplar material was mentioned in other contexts such as having work on display and asking students to mark work using a set of criteria provided by the teacher.

Almost all the teachers mentioned some form of self-assessment in their plans, ranging from using red, amber or green 'traffic lights' to indicate the student's perception of the extent to which a topic or lesson had been understood, to strategies that encouraged self-assessment via targets which placed responsibility on students (e.g., 'One of these twenty answers is wrong: find it and fix it!'). Traffic lights (or smiley faces—an equivalent that did not require coloured pens or pencils!) were seen in about half of the plans and in practically all cases their use was combined with strategies to follow up the cases where the students signalled incomplete understanding. Several teachers mentioned their conviction that group work provided important reinforcement for students, as well as providing the teacher with insights into their students' understanding of the work.

We were interested in whether the choices of activities by the different teachers showed any structure (e.g. do particular combinations of strategies occur together?). However, use of cluster analysis and multidimensional scaling (Schiffman *et al.*, 1981) revealed no tendency for particular strategies to be found together. In this

sense, the strategies and techniques appear to be relatively independent of one another.

The other component of the intervention, the visits to the schools, provided an opportunity for project staff to discuss with the teachers what they were doing, and how this related to their efforts to put their action plans into practice. The interactions were not directive, but more like a holding up of a mirror to the teachers. Since project staff were frequently seen as 'experts' in either mathematics or science education, there was a tendency sometimes for teachers to invest questions from a member of the project team with a particular significance, and for this reason, these discussions were often more effective when science teachers were observed by mathematics specialists, and vice-versa.

We aimed for each teacher to be observed at least once each half term, although releasing teachers to discuss their lessons either before or afterwards was occasionally a problem (and schools that had guaranteed teacher release for this purpose at the beginning of the project were sometimes unable to provide for it).

A detailed description of the qualitative changes in teachers' practices is beyond the scope of this paper (see Black *et al.*, 2003, for a full account), but it is worth noting here that the teachers' practices were slow to change, and that most of the changes in practice that we observed occurred towards the end of the year, so that the actual size of the effects found are likely to be underestimates of what could be achieved when teachers are emphasizing formative assessment as an integral part of their practice.

Research design

Given the nature of the intervention, which was designed to build on the professionalism of teachers (rather than imposing a model of 'good formative assessment' on them), we felt that to utilize a traditional research design on the teachers would have been inconsistent. Furthermore, it would have been impractical. Since each teacher was free to choose which class would be the focus for this work, there was no possibility of standardizing either the 'input' or 'output' variables. For this reason, the collection of empirical quantitative data on the size of effects was based on an approach which we have termed 'local design'. Drawing more on interpretivist than positivist paradigms, we sought to make use of whatever assessment instruments would have been administered by the school in the normal course of events. In many cases, these were the results on the national tests for 14-year-olds or the grades on the national school-leaving examination (the GCSE), but in some cases we made use of scores from school assessments (particularly in science, where modular approaches meant that scores on end-of-module tests were available).

Using externally mandated tests and examinations as 'input' and 'output' variables has both weaknesses and strengths. On the minus side, such tests might lack curricular validity (McClung, 1978) in that they may not accurately reflect what the teachers were teaching in their classrooms. On the other hand, to require teachers to develop additional assessments specifically related to what they had been teaching would have been an unacceptable addition to their already heavy workloads. Nor

would providing our own assessments have been a satisfactory solution, since this would immediately raise questions of whether they captured what the teachers were trying to achieve. Furthermore, all the teachers were happy with the ‘output’ variables we had suggested as satisfactory measures of what they *were* trying to achieve in their classrooms, suggesting a considerable degree of ‘alignment’ between their teaching and the assessments used (although it is worth noting here that the teachers were critical of these assessments, because they felt that the assessments did not assess the important aspects of the subject). While the use of external tests therefore raises many issues, we do not think that any other approach would have been appropriate.

For each focal class we therefore had a focal variable (that is, dependent variable or ‘output’) and, in all but a few cases, we also had reference variables (that is, independent variables or ‘inputs’). In order to be able to interpret the outcomes we discussed the local circumstances in the school with each teacher and set up the best possible comparison group consistent with not disrupting the work of the school. In some cases this was a parallel class taught by the same teacher in previous years (and in one case in the same year). In other cases, we used a parallel class taught by a different teacher and, failing that, a non-parallel class taught by the same or different teacher. We also made use of national norms where these were available. In almost all cases, we were able to condition the focal variable on one or more reference variables, although in some cases the reference variables were measures of general ability (e.g. the National Foundation for Educational Research—NFER’s Cognitive Abilities Test) while in others they were measures of achievement in that subject (e.g. end-of-year-8 tests).

In order to be able to compare the results, raw differences between experimental and comparison groups were standardized by dividing by the pooled standard deviation of the experimental and comparison scores. This measure, called either the ‘standardized effect size’, or sometimes just ‘effect size’, and denoted d provides a way of comparing experimental results achieved with different measures (although see the discussion of problems in interpreting effect sizes below).

Results

Of the 24 teachers originally selected, 22 remained part of the project until its conclusion in July 2000. Peter from Brownfields School formally withdrew from the project and Lisa left Riverside School, to be replaced by Patrick. However, several teachers left their schools at the end of the project, and reliable data were available for only 19 teachers, four of whom had decided to have two focal classes each, resulting in data on 23 classes. For two of the classes (Nancy and James) there were two possible comparison groups. In the case of James, the effects are comparable ($d = 0.29$ and $d = 0.38$). However, in the case of Nancy, comparison with another teacher (actually Patrick, who was not originally part of the study) yields a negative effect ($d = -0.31$) while a comparison with a similar set taught in the previous year by Nancy yields a very large positive effect ($d = 1.15$). For reasons of completeness, both results for Nancy and James have been included, giving a total of 25 effect sizes, which are shown in Table 4, and summarized in stem-and-leaf form in Figure 1.

Table 4. Experimental results for the 25 teachers involved in KMOFAP

School	Subj	Teacher	Yr	Set	n	Focal variable	Reference variables	Comparison group	n	SD	Raw effect	d	p
BF	M	Iwan	7	1	25	SE7	C7	D	95	17.54	+ 6.63	+ 0.38	0.0299
BF	M	Iwan	9	1	27	KS3	C7, S8	D	94	33.84	12.25	+ 0.36	0.0081
BF	M	Lily	7	3	25	SE7	C7	D	95	14.96	- 5.22	- 0.35	0.1434
BF	S	Rose	7	5	8	SE7	SB7	D	25	24.80	38.44	+ 1.55	0.0001
BF	S	Peter											
CE	M	Belinda	8	1	21	SE8	KS2	P	26	10.61	2.76	+ 0.26	0.3604
CE	M	Angela	9	3	23	KS3	KS2	D	21	15.93	19.12	+ 1.20	0.0001
CE	S	Sian	8	-	26	SE8	SE7	P	169	0.889	0.342	+ 0.38	0.0113
CE	S	Carl	8	-	27	SE8	SE7	P	169	0.911	0.417	+ 0.46	0.0018
CI	S	Derek	9	2	27	KS3	C7	D	56	0.666	0.183	+ 0.27	0.1984
CI	S	Philip	9	1	29	S3	C7	P	56	0.695	0.169	+ 0.24	0.2305
CI	M	Greg	9	4	24	KS3	SE7	P	20	0.0379	- 0.025	- 0.07	0.8045
CI	M	Eva	9	1	29	KS3	SE7	P	28	0.4916	- 0.127	- 0.26	0.3997
RS	M	Nancy	8	1	32	KS3	C7	P*	34	38.7	- 12	- 0.31	0.0019
RS	M	Nancy	8	1	32	KS3	C7	S	30	27.8	+ 32	+ 1.15	0.0001
RS	M	Nancy	9	1	34	KS3	KS2	N	-	0.50	0.13	+ 0.26	0.0669
RS	M	Patrick	9	1	30	KS3	KS2	N	-	0.58	0.38	+ 0.66	0.0001
RS	M	Lisa											
RS	S	Jerry	8	2									
RS	S	Tom	8	2	32	SE8	-	P	34	43.38	+ 10.02	+ 0.23	0.3852
TB	S	James	11	1	32	GCSE	-	S	32	0.879	0.255	+ 0.29	0.2628
TB	S	James	11	1	32	GCSE	-	P	32	1.013	0.375	+ 0.38	0.1038
TB	S	Robert	9	-	30	KS3	SE8	I	56	15.33	2.95	+ 0.19	0.1438

TB	M	Steve	11	2	32	GCSE	KS3	P	31	0.941	0.380	+ 0.40	0.1093
TB	M	Steve	11	4	24	GCSE	KS3	D	87	1.48	0.222	+ 0.15	0.2849
TB	M	Kerry	11	4	23	GCSE	KS3	D	87	1.54	0.309	+ 0.20	0.1348
TB	M	Kerry	11	1	32	GCSE	KS3	D	82	1.95	0.4786	+ 0.25	0.0276
WF	M	Gwen	9	2	23	KS3	—	L	24	0.462	0.158	+ 0.34	0.2469
WF	M	Alice											
WF	S	Susan											
WF	S	Kieron											

Key**Focal variables**

KS3 Key Stage 3 tests

SBn School-produced test at beginning of year n

SEn School-produced test at end of year n

Reference variables

Cn CAT scores in year n

SEn School produced tests at end of year n

Comparisons

I Parallel set taught by same teacher in same year

S Similar set taught by same teacher in previous year

P Parallel set taught by different teacher in same year

L Similar set taught by different teacher in previous year

D Non-parallel set taught by different teacher in same year

N National norms

★ Non-representative comparison

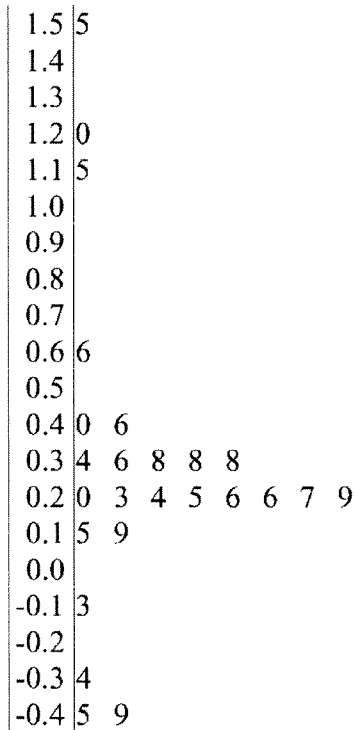


Figure 1. Overall standardised effect sizes

NB: in this stem-and-leaf diagram, negative values are displayed following the convention that 'data = model + residual' so that a value of -0.26 is shown with a 'stem' of -0.3 and a 'leaf' of 4 (representing 0.04).

As can be seen, the majority of effect sizes are around 0.2 to 0.3, with a median value of 0.27. Given the fact that each of these results is a separate, 'mini-experiment', care needs to be taken in drawing any general conclusions about the net effect of the adoption of formative assessment (see 'Discussion' below). The mean effect size is 0.34, but is clearly influenced by some extreme values, and since the effect sizes are not normally distributed, the jack-knife procedure recommended by Mosteller and Tukey (1977) was used to provide an estimate of the true mean effect as 0.32 and a 95% confidence interval of the true effect size as (0.16, 0.48).

In order to examine the relationship between a teacher's practice and the effect sizes, we classified teachers into one of four groups, according to their use of formative assessment strategies in their classrooms, as shown in Figure 2.

These characterisations had emerged from our observations of each teacher's practice, and were based on their use of key strategies during the main period of the project. Independent classification of the 24 teachers by two of us (CH and CL) produced identical classification for all but two teachers, and these were resolved after discussion. These classifications were produced before the results were known. The effect sizes by teacher type are shown in Table 5. Although there is no obvious trend in terms of average effect size, as one moves from less to more expert teachers,

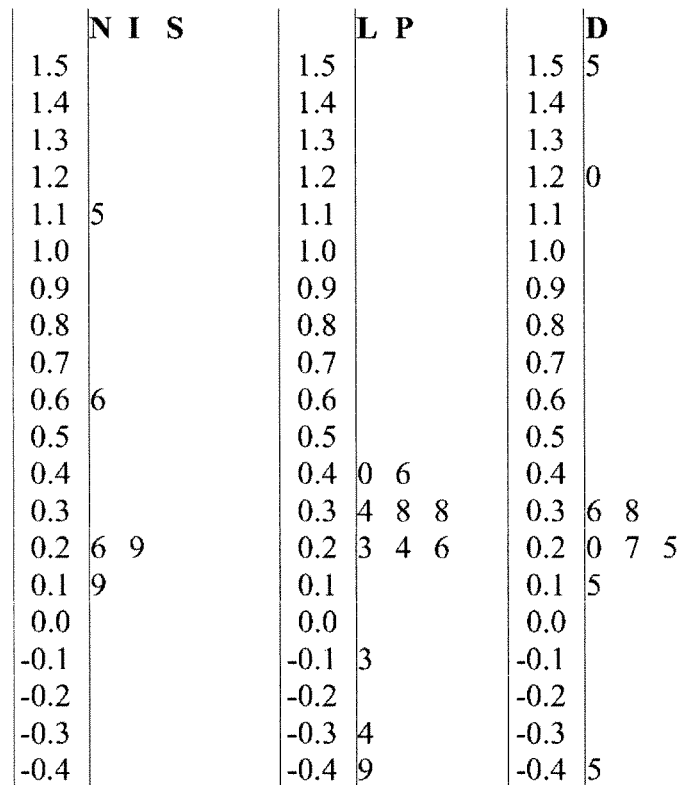


Figure 2: Standardized effect sizes by comparison type

the interquartile range of effect sizes reduces, indicating further support for the attribution of the effects to the quality of formative assessment.

An analysis of the effects by different forms of comparison group in the form of side-by-side stem-and-leaf diagrams (Figure 2) shows that no significant difference in effect sizes for the different form of comparisons is apparent.

There was no difference in the mean effect size for groups of different ages,

Table 5. Effect sizes classified by teachers' use of formative assessment strategies

Group	Count	Median	H-spread*
Experts	7	0.25	0.07
Moving pioneers	10	0.31	0.25
Static pioneers	2	1.38	0.35
Triallers	6	0.15	0.64

*The 'hinge-spread' or 'H-spread' is analogous to the interquartile range, and tends to it as the number of data-points increases, but is more easily interpreted for small samples. See Tukey (1977) for details.

although it is worth pointing out that the year 11 focal groups—where the ‘output’ measure was the grade on the GCSE examination—all had positive effect sizes. There was no systematic variation in effect size with the ability of the focal class although an analysis by subject shows that all the negative effect sizes were found for the mathematics groups (the median effect sizes for the mathematics and science groups were, however, almost identical).

Discussion

By its very nature, the quantitative evidence provided here is difficult to interpret. The comparisons are not equally robust. In some cases, we have comparisons with the same teacher teaching a parallel class in previous years, which, in terms of the main question (that is, has the intervention had an effect?) is probably the best form of comparison. In other cases, we have comparisons with a different teacher teaching a parallel set, so it could be that in some cases a positive effect indicates only that the teacher participating in the project is a better teacher than the teacher teaching the comparison class. In other cases, the comparison class is another class (and sometimes a parallel class) taught by the same teacher, and while there are examples of positive effect sizes here (in the case of Robert, for example) it is also reasonable to assume that the observed size of such effects will be attenuated by what we have termed ‘uncontrolled dissemination’. Indeed, while two of our teachers did view involvement in the project as a short-term commitment (after which they would return to teaching ‘normally’) for the vast majority of our teachers, involvement in the project has not just spread to all their classes, but has fundamentally altered their views of themselves as professionals. In some cases, the only comparisons available were the classes of different teachers teaching non-parallel classes, and given the prevalence of ability grouping in mathematics and science, and its effect on achievement (see Wiliam & Bartholomew, 2004), disentangling the effect of our interventions from contextual factors is quite impossible.

In particular, the problematic nature of the comparison groups makes interpretation of the four negative effects difficult. The case of Nancy has been discussed above. For Lily, an inexperienced female teacher teaching in an all-boys school, the only comparisons possible were with more experienced, male teachers, and therefore a negative effect is not surprising (and although this may seem like special pleading, it is our belief, from observations of her teaching, that Lily did improve substantially in her teaching during the project). We could also attempt to ‘explain away’ the negative effects for the two mathematics teachers at Century Island School by citing their limited attendance at the in-service sessions. Such exclusions would be warranted if we were seeking to establish the effectiveness of formative assessment, since their engagement with formative assessment was, in fact, very limited. However, since our focus in this paper is not whether formative assessment is effective in raising achievement (because there is significant research in existence to show that it is), but on how to support teachers in developing their practice, we believe that it is appropriate to include these results.

Although a variety of measures were used as inputs and outputs, the very fact that

these were either national tests and examinations, or assessments put in place by the school, gives us a degree of confidence that these measures have some validity in terms of what the teachers were trying to achieve. There is also a problem inherent in the standard definition of standardized effect size in the literature (Glass *et al.*, 1981 for example). While standardized effect sizes are more comparable than raw scores, and allow different assessments to be placed on a common metric, there are nevertheless significant problems of interpretation. Dividing the difference between comparison and experimental groups by the pooled standard deviation clearly makes sense in that an improvement of 5 marks from (say) 60 to 65 represents a huge improvement if the standard deviation is 5 marks, but only a modest improvement if the standard deviation is 20 marks. However, the same logic dictates that an average improvement of half a grade per student at GCSE is to be regarded as a bigger effect if it is achieved in a top set where the standard deviation is one grade ($d = 0.5$), than if it is achieved in a mixed-ability class where the standard deviation is two GCSE grades ($d = 0.25$).

There is also the question of what lessons can be drawn from these six schools for policy and practice elsewhere. After all, the schools in this study were not typical, in that they had identified themselves as interested in exploring the development of formative assessment, and they were given a much greater degree of support than is available to most teachers. In response, we can only agree, and repeat the usual researchers' litany that 'more research needs to be done'. However, in defence of the idea that further research in this area is worth undertaking, we would make two points.

The first is that while the schools were a selective sample, the teachers were much less so, representing a range of expertise and experience, and almost all the teachers appear to have improved. Furthermore, these teachers have generated a series of 'living examples of implementation' that have served to make it easier to introduce these ideas to other teachers (Black *et al.*, 2002). The experience of the participating teachers is already being built upon in their schools and Local Education Authorities (and more broadly—see Black & Wiliam, 2003), but of course it remains to be seen to what extent this work can be scaled up to an LEA or a country.

The second relates to the cost of the support. We estimate that the cost of providing the support (as opposed to researching its effects) was around £2000 (\$3000) per teacher or approximately 8% of the salary costs for one teacher for one year. While this is much more than most schools have available per teacher for professional development, it is a relatively small proportion of the annual cost of each teacher (especially if, as appears to be the case, this is a one-off, rather than a recurrent cost).

In conclusion, despite the cautions noted above, we believe that the results presented here provide firm evidence that improving formative assessment does produce tangible benefits in terms of externally mandated assessments (such as key stage 3 tests and GCSE examinations in England). Placing a quantitative estimate on the size of the effect is difficult but it seems likely that improvements equivalent to approximately one-half of a GCSE grade per student per subject are achievable. While these improvements might sound small, if replicated across a whole school,

they would raise the performance of a school at the 25th percentile of achievement nationally into the upper half. At the very least, these data suggest that teachers do not, as is sometimes reported, have to choose between teaching well and getting good results.

Notes on contributors

Dylan Wiliam took up the post of Director of the Learning and Teaching Research Center at ETS in September 2003. Before that he was Professor of Educational Assessment at King's College London, which he joined after teaching mathematics and science in secondary schools in London. At King's, he worked on various research projects related to assessment, including Graded Assessment in Mathematics, the Consortium for Assessment and Testing in Schools, as well as teaching on PGCE, Masters, Ed.D. and Ph.D. programmes.

Paul Black is Emeritus Professor at King's College London. During his career he has been involved in a range of Nuffield curriculum projects and in many research projects, mainly in science education and assessment. In 1987–8 he chaired the task group (TGAT) that advised ministers on the new national assessment and testing policy. Since his retirement he has concentrated on the study of formative assessment.

Chris Harrison taught science in several schools in the London area before she joined King's College in 1993. She now spends part of her time working with trainee teachers. Her work in formative assessment at King's has led to several research and professional development projects in the UK and abroad at both primary and secondary level. She has also worked as Assessment for Learning consultant on national projects in England, Scotland and Jersey.

Clare Lee was the research fellow on the KMOFAP project at King's, and had previously taught secondary mathematics for over twenty years in several schools. She has also worked with teachers on Action Research projects based at Oxford University. Currently she is Teacher Advisor for Assessment for Warwickshire working on many Assessment for Learning projects at primary and secondary level within the LEA.

Note

This is a revised version of a paper presented at the 27th annual conference of the British Educational Research Association, University of Leeds, September 2001. In revising the paper, we have taken into account the comments of two anonymous reviewers, whose help we gratefully acknowledge, although we are entirely responsible for any remaining errors.

References

- Black, P., Harrison, C., Lee, C., Marshall, B. & Wiliam, D. (2002) *Working inside the black box* (London, King's College London Department of Education and Professional Studies).
- Black, P., Harrison, C., Lee, C., Marshall, B. & Wiliam, D. (2003) *Assessment for learning: putting it into practice* (Buckingham, Open University Press).

- Black, P. & Wiliam, D. (2003) In praise of educational research: formative assessment, *British Educational Research Journal*, 29(5), 623–637.
- Black, P. J. & Atkin, J. M. (Eds) (1996) *Changing the subject: innovations in science, mathematics and technology education* (London, Routledge).
- Black, P. J. & Wiliam, D. (1998a) Assessment and classroom learning, *Assessment in Education: principles, policy & practice*, 5(1), 7–73.
- Black, P. J. & Wiliam, D. (1998b) *Inside the black box: raising standards through classroom assessment* (London, King's College London School of Education).
- Boaler, J. (2002) *Experiencing school mathematics: traditional and reform approaches to teaching and their impact on student Learning* (Mahwah, NJ, Lawrence Erlbaum Associates).
- Brousseau, G. (1984) The crucial role of the didactical contract in the analysis and construction of situations in teaching and learning mathematics, in: H-G., Steiner (Ed.) *Theory of mathematics education: ICME 5 topic area and miniconference* (Bielefeld, Germany, Institut für Didaktik der Mathematik der Universität Bielefeld), 110–119.
- Brown, M. L. (Ed.) (1988) *Graded Assessment in Mathematics development pack: teacher's handbook* (Basingstoke, Macmillan).
- Crooks, T. J. (1988) The impact of classroom evaluation practices on students, *Review of Educational Research*, 58(4), 438–481.
- Glass, G. V., McGaw, B. & Smith, M. (1981) *Meta-analysis in social research* (Beverly Hills, CA, Sage).
- McClung, M. S. (1978) Are competency testing programs fair? Legal? *Phi Delta Kappan*, 59(6), 397–400.
- Mosteller, F. W. & Tukey, J. W. (1977) *Data analysis and regression: a second course in statistics* (Reading, MA, Addison-Wesley).
- Natriello, G. (1987) The impact of evaluation processes on students, *Educational Psychologist*, 22(2), 155–175.
- Newmann, F. M., Bryk, A. S. & Nagaoka, J. K. (2001) *Authentic intellectual work and standardized tests: conflict or coexistence?* (Chicago, IL, Consortium on Chicago School Research).
- Nuthall, G. & Alton-Lee, A. (1995) Assessing classroom learning: how students use their knowledge and experience to answer classroom achievement test questions in science and social studies, *American Educational Research Journal*, 32(1), 185–223.
- Paris, S. G., Lawton, T., Turner, J. & Roth, J. (1991) A developmental perspective on standardised achievement testing, *Educational Researcher*, 20(5), 12–20.
- Schiffman, S. S., Reynolds, M. L. & Young, F. W. (1981) *Introduction to multidimensional scaling: theory, methods, and applications* (New York, Academic Press).
- Tukey, J. W. (1977) *Exploratory data analysis* (Reading, MA, Addison-Wesley).
- Wiliam, D. (2003) The impact of educational research on mathematics education, in: A. Bishop, M. A. Clements, C. Keitel, J. Kilpatrick & F. K. S. Leung (Eds) *Second international handbook of mathematics education* (Dordrecht, Netherlands, Kluwer Academic Publishers).
- Wiliam, D. & Bartholomew, H. (2004) It's not which school but which set you're in that matters: the influence of ability-grouping practices on student progress in mathematics, *British Educational Research Journal*, 30(2), 279–293.